DBS
Dublin Business School
excellence through learning

QQI

Data and Web Mining

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

APRIL 2017 FT GROUP BLOCK 3 RESIT

Module Code: **B8IT108**

Module Description: **Data and Web Mining**

Examiner: **Terri Hoare**

Internal Moderator: **Darren Redmond**

External Examiner: **Dr Brett Becker**

Date: 08/01/18
Time: 10:00-12:00

INSTRUCTIONS TO CANDIDATES

Time allowed is 2 hours
Question 1 is mandatory (30 marks)
Answer 2 of the remaining 3 Questions (35 marks each)

All answers should reference literature and case studies as appropriate. Use of scientific calculators only is permitted.

Question 1 : MANDATORY **(30 Marks)**

a) Use the AllElectronics Customer Database Table below to build a Naïve Bayes Classifier that can be applied to predict the class label buys_computer for the unlabelled tuple X given below.

X = (age=youth, income=medium, student=yes, credit_rating=fair)

Hint: apply the Naïve Bayesian formula

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} where \;\; P(X|C_i) = P(x_1|C_i) * P(x_2|C_i) * \ldots * P(x_n|C_i)$$

Electronics Customer DB - Training Tuples - Label buys_computer

RID	age	income	student	credit_rating	buys_computer
1	youth	high	No	fair	**NO**
2	youth	high	No	excellent	**NO**
3	mid-aged	high	No	fair	**YES**
4	senior	medium	No	fair	**YES**
5	senior	low	Yes	fair	**YES**
6	senior	low	Yes	excellent	**YES**
7	middle-aged	low	Yes	excellent	**YES**
8	youth	medium	No	fair	**NO**
9	youth	low	Yes	fair	**YES**
10	senior	medium	Yes	fair	**YES**
11	youth	medium	Yes	excellent	**YES**
12	middle-aged	medium	No	excellent	**YES**
13	middle-aged	high	Yes	fair	**YES**
14	senior	medium	No	excellent	**NO**

[10]

b) Appraise the role of the confusion matrix in measuring the performance of a trained classifier.

[10]

c) Propose and evaluate two metrics commonly used for measuring the performance of clustering algorithms.

[5]

d) Explain what is meant by the metric TF-IDF in text mining.

[5]

ANSWER (2 OUT OF 3) QUESTIONS BELOW

Question 2 **(35 Marks)**

a) Appraise the use of the Decision Tree Algorithm in supervised learning under the headings of model description, model parameters, inputs, outputs, pros, and cons.

[15]

b) Propose two applications for deployment of a decision tree classifier.

[6]

c) What is the strong 'naïve' assumption made by a Naïve Bayes Classifier?

[2]

d) Propose a technique for minimising the effects of this assumption.

[2]

e) Ensemble Models are used in many practical classification problems.

Propose two techniques for use in ensemble modelling that improve the overall error rate and reduce the bias of individual models.

[10]

Question 3 **(35 Marks)**

a) Appraise the main differences between Classification and Clustering algorithms and propose an application where clustering is used in the pre-processing step for use by a Classification algorithm. [10]

b) Appraise the use of k-Means as a clustering algorithm considering how the algorithm works, input and output limitations, pros, cons, parameters, and types of suitable applications. [10]

c) Propose a use case where DBSCAN (Density-Based Spatial Clustering of Application with Noise) for clustering would be preferable to k-Means clustering.

[4]

d) Explain what is meant by a SOM (Self Organising Map) and propose a use case for a SOM in data mining.

[5]

e) Identify differences between the Apriori and FP-Growth algorithms used for mining frequent itemsets when building Recommender Engines.

[6]

Question 4 **(35 Marks)**

a) Cluster analysis of data represented in graphs and networks extracts valuable knowledge and information. Propose two applications where graph clustering is beneficial outlining both the graph structure and the benefits gained through clustering the data.

[10]

b) Propose a text mining technique for clustering documents of unstructured text.

[10]

c) PageRank is an algorithm used by Google Search to rank websites. Describe the measure of network centrality that it uses and contrast with at least two other measures of 'centrality' or power in a network graph.

[15]

END OF EXAMINATION